Activity Recognition from Video Data using Spatial and Temporal Features

Mohamad Al-Wattar, Rinat Khusainov, Djamel Azzi, and John Chiverton
School of Engineering
University of Portsmouth
Portsmouth, United Kingdom
{mohamad.alwattar, rinat.khusainov} @port.ac.uk

***Abstract*—A method to monitor elderly people in an indoor environment using conventional cameras is presented. The method can be used to identify people's activities and initiate suitable actions as needed. The originality of our approach is in combining spatial and temporal contexts with the position and orientation for the detected person. Preliminary evaluation, based only on the first two features (spatial and temporal), achieved the accuracy over 60% in a realistic residential environment. Although the results are based on using only two out of the four proposed input features, they already demonstrate a promising improvement over using a single feature in isolation**.

Keywords—Assisted living, activity recognition in video, temporal and spatial contexts, context fusion

I. INTRODUCTION

The growing aging population increases dramatically the load on healthcare and social services, making traditional care models relying on care and medical staff, unsustainable in the long term [1, 2]. The main idea behind Assisted Living (AL) is to use technological solutions to help people live independently, while maintaining their quality of life and care. Our focus is on helping elderly to live (and work) independently in their own home environment. Many companies and researchers are currently pursuing this objective [3, 4]. Some of them used attached hardware and wearable devices, like wristbands or pendants. Others focused on providing help using telehealth solutions.

Using contactless sensors, especially cameras, for assisted living is still in its infancy, but it is a very promising solution. Such contactless approaches can be cheap, easy to install, and more scalable compared to other systems. The main objective of our work is to develop a method to monitor elderly in an indoor environment using conventional cameras, identify people's activities, and initiate suitable action as needed.

Most of the previous work in this area (as described in Sec. 2) relied on using only a small number of features in video data. For instance, some system only used the person's location (spatial data) or position (posture).

The originality of our approach is in combining spatial and temporal contexts with the position and orientation for the detected person. The intuition behind this idea is that using additional contexts can help to disambiguate between different activities that may be happening at the same location and to deal with noise in data and errors in video and image processing.

The main contributions of this paper are as follows:

- A conceptual design of the proposed method.
- A prototype implementation that can recognise a variety of activities of daily living.
- Preliminary evaluation in a realistic residential environment that demonstrates the potential benefits of the proposed method.

II. RELATED WORK

There has been a significant amount of prior work done on techniques for identifying human activities. Smart cameras were used to identify activities of elderly people at home [5]. They managed to detect and identify humans based on colour using colour conversion from RGB to HSV. This study also used a colour system for camera handover. An SVM-based algorithm was then used to predict activities, including falls detection. This study has a number of issues. It only detects one person. It is expensive and has some limitation in detecting human bodies, because it is based on colour. SVM and Hidden Markov Models (HMMs) were used with smart cameras for activities recognition and tracking [6]. The cameras can identify a fall when it happens. However, it is also an expensive solution.

Several studies suggested the use of the Microsoft's Kinect sensor to detect, track, and identify human activities. For instance, a robot with an attached Kinect sensor was used to follow a person [7]. The system detects falls by dividing the human body into parts, and sends an emergency SMS when a fall happen. A drawback of this technique is that it requires a robot with a Kinect sensor to follow the person at all times. Also, the distance from the Kinect sensor affects the accuracy of the technique. Kinect depth sensors and skeleton tracking were used for person localisation [8]. In this example, two Kinects work together. The first one works as a voice recognition and localisation sensor, and the second Kinect detects the skeleton. The introduction of the Kinect v2 allowed researchers to use it for predicting heart attacks [9]. A message can be sent or a video conference can be started with a doctor or nurse, based on the heart rate monitoring system that Kinect v2 has. However, the patient must be facing the Kinect sensor within a fixed distance to allow the sensor to measure the heart rate. A webcam was combined with an IR sensor and Kinect sensors to identify activities [10]. The approach was able to detect falls and fall related activities based on fuzzy clustering. A Kinect sensor was used for activities segmentation and background subtraction, with a mixture

of Gaussian used for detection. The resulting silhouette from the detection was supplied to the activity segmentation system. This study identifies only small number activities: sitting, standing, and falling.

Two fisheye cameras and IR lights were used in an automatic monitoring system for elderly people [11]. The system detects the person based on silhouette extraction with the use of spatio-temporal filtering to reduce the noise in the detected image. To improve the results, a median filter is used, and tracking of the silhouette is done with a Kalman filter. SVM classifier is used to identify the action. According to this study, seven human actions were identified: lying on bed, standing, walking, sitting, falling, sitting on the bed, and falling from the bed. The major drawback in this work is that it uses separate IR lights that need to be installed on a standing lamp, and it only works with one person.

The person silhouette was also used to recognise people's activities [12]. This work is based on background subtraction and colour enhancement to extract the silhouette, and it uses an SVM classifier. However, this approach works with only one person at a time, and does not have the tracking ability.

The W4 system uses a normal camera that works with black and white images [13]. First, the system scans the background and identifies new objects in the images based on foreground subtraction. After that, it can distinguish humans from other objects by using shapes and periodic motion cues. W4 can identify actions and differentiate between humans based on head detection and person segmentation, posture analysis, and body part detection. The tracking operation works by matching the template and predicting the motion, employing tracking algorithm to estimate the position of the torso for each human. The tracking is based on silhouette and body part detection. W4 was able to achieve a good accuracy in tracking, detecting and recognising some actions (e.g. standing, bending, lying, and sitting) based on information gathered from 170 silhouettes. Also, it can identify whether the person is carrying an object in their hands. To detect a face, W4 combines two methods: shape cues and vertical projection. W4 is a single camera system that works with only black and white images, and in an outdoor environment.

Motion sensors have been employed to identify human activities based on spatial information to improve performance [14]. Also, sensors have been used to recognise an activity based on temporal information. In particular, time of the day, day of the week, and seasons were used to improve the accuracy [15]. Neural network and images from two orthogonal cameras were used to identify human activities based on two feature vectors [16].

III. METHOD

This section describes a method that can identify people's activities inside a house. The main idea behind our approach is to combine several spatial and temporal features for recognising activities of daily living. The location of a person (a spatial feature) can give information about the person's current activity. For instance, if the person is in front of a cooker, then they may be cooking. However, relying on the location alone may not be sufficient to correctly recognise the activity. An activity can be ongoing, but location can be detected incorrectly for a short time, or the actual location can change for a short time (for instance, when the person steps away and then quickly comes back to what they were doing).

The location can be detected incorrectly at the start or at the end of an activity due to fuzzy boundaries between objects. A new activity can start, but the person can be still at a pervious location (for instance, when the two locations partially overlap). Similarly, the location can change, but the previous activity can still be continuing.

Using temporal features can help address some of these problems. In particular, the system can monitor how long the person has spent at a particular location and use that information in combination with the location data to recognise the activity. The idea here is that if a person has spent a longer time at the location, then they are more likely to be engaged in the corresponding activity, rather than just passing by.

Using the person's orientation and posture can further improve the accuracy of activity recognition, particularly when changing from one activity to another. This is likely to be helpful in addressing with overlapping locations. Finally, performance can be improved by tailoring the system parameters for the specific activities and the particular person. Fig. 1 shows how these different inputs are combined in our proposed approach**.**

The first step for the proposed method is to detect the person in a live video stream. The system also needs to identify locations of different objects in the room, such as furniture or appliances. In a simple case, object locations can be specified manually for each room (see Fig. 1).

In a more advanced scenario, the system can identify objects automatically from the live video using such techniques as deep learning [17]. The coordinates of the person are compared to the coordinates of room objects to identify the person's location. The system also determines the person's position (posture) and orientation from the video. The resulting four inputs are then combined to recognise the current activity.

This paper focuses on using two of the proposed four features for identifying activities: the current location and the time spent at that location.

A simple way to use time to differentiate between engaging in the activity at a particular location and just passing by is using a time threshold. A new activity is detected only when the person spends at a new location longer than the time threshold. For example, if the person is cooking and then goes to a dining table to bring a plate, the detected location may change, but the actual activity of cooking will continue. Adding a time threshold for changing the activity following a change in location will help in such cases, as it will not change the activity

immediately after the new location is detected. This can, however, introduce new errors when switching from one activity to another. Introducing a time threshold will cause delays when detecting a new activity. It may also result in missing short activities, if their duration falls below the threshold. Therefore, selecting the threshold value can have a significant effect on the system performance.

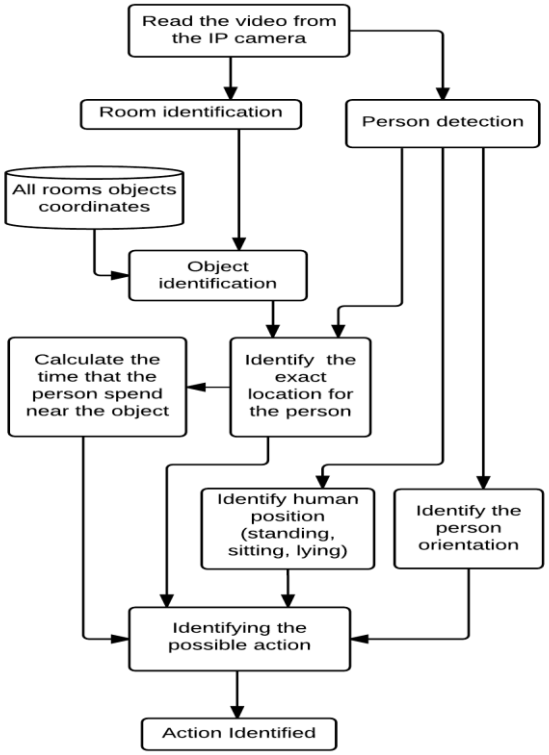

Figure 1. Overall method design.

The goal would be to achieve a balance between setting too low threshold resulting in detecting spurious activities, and too high threshold making the system less responsive and missing short activities. This suggests that an activity specific threshold can be better.

To calculate activity-specific thresholds, it is proposed to analyse available video data to identify the instances when the location changes for a limited period, but the activity continues. The time threshold for each activity can then be calculated as the minimum, maximum, or average duration of such instances.

Several activities can share similar temporal patters. Therefore, clustering activities can be used when dealing with larger numbers of activities. In that case, cluster specific time thresholds can be used during activity recognition. A number of activity properties can be used to do clustering. These can include the activity frequency, room location (in a house), and activity duration. For example, activities in the kitchen may tend to be shorter (requiring a lower time threshold) than activities in the bedroom. Infrequent activities, such as using a washing machine, may need a higher threshold, as it is more likely that the person is just passing by a washing machine without the intention to use it. Activities that normally take longer durations, such as cooking, may similarly require a higher threshold compared to activities that are usually short, such as making a cup of tea. The next section presents preliminary results comparing the different approaches discussed above

IV. Results

To evaluate the feasibility of the proposed approach, preliminary experiments were carried out using a specially designed test facility. The facility, called Port-Eco house, is a normal living environment, where people can live and carry out their normal daily activities. The house has the appropriate infrastructure that allows researchers to install the cameras wherever is required. In addition to that the Port-Eco house has a secure connection to cameras and computers via a VPN and/or a Secure Shell SSH.

Video recordings of everyday activities were made in two locations: the kitchen and the home office. Each video was manually labelled to mark the start and end times of different activities in the video. Also, locations of different objects in the video (such as cooker, sink, kettle, computer, etc.) were manually defined for each video.

Background subtraction with KNN was used to detect the current location of the person in videos. This technique detects any movements even due to a change in lighting. Erode and dilate morphology operators were employed to discard the unnecessary noise in the video due to such false movements. To further address the noise issue, only objects with an area larger than 4000 pixels were used for detection, because they would be more likely to correspond to a human in the picture. The method described in Sec. 3 was then applied to the detected sequence of locations to recognise the activities. Fig. 2 shows a sample output from our system.

Figure 2. Output of the activity recognition system.

The screenshot in Fig. 2 shows the detected person surrounded by the blue rectangle, the instant detected location, the last detected location, and the time that the person spent in the location. The recognized activity is also shown in the video.

In the kitchen, the system identified 75 instances of changing the person's location. The accuracy of identifying the correct activity in each location instance was calculated to evaluate the effectiveness of our method.

In our experiments, the method performance was analysed for different activity independent time threshold values, and compared with activity specific time thresholds.

The activity independent time threshold values ranged from 0 to 252 frames, which was the longest activity duration at any location. Three methods were used for the activity specific thresholds: minimum, maximum, and average (activity duration based on the test videos), as described in Sec. 3. The resulting accuracy for each case is presented in Fig. 3. In this work, the accuracy calculated by counting the total number of correctly identified activities divided by the total number of incidents.

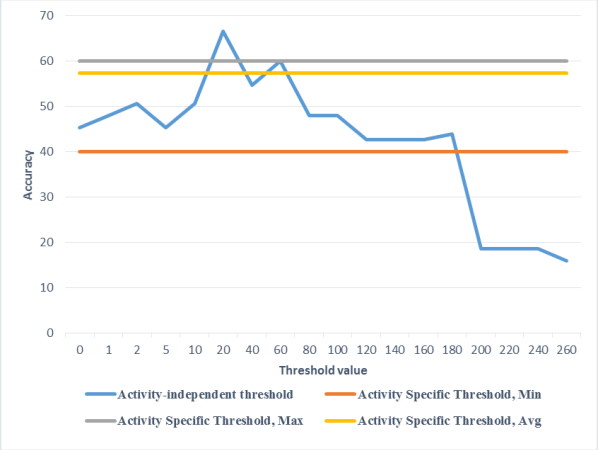

Figure 3. The system performance for different time thresholds.

While the activity independent threshold outperforms activity specific thresholds in the best-case scenario (20 for our data), there is a large variation in accuracy depending on the threshold value. This means that the system performance would be very sensitive to the correct threshold calibration. Activity specific threshold shows a more robust performance, particularly when the maximum duration is used (see Sec. 3). We believe that this advantage will become more evident when more video data are used to calculate the thresholds and to test the accuracy. The activity specific threshold based on average duration also shows a good performance and is likely to help dealing with extreme cases, when more video data is available. Most importantly, combining time and location (i.e. using a time threshold) gives a better accuracy than using location alone (i.e. zero threshold in Fig. 3).

V. Conclusion

A novel approach to identifying activities of daily living using conventional cameras was described. The approach is based on combining spatial and temporal contexts with the position and orientation for the detected person. Although the presented preliminary results are based on using only two out of the four proposed input features (spatial and temporal), they already demonstrate a promising improvement over using a single feature in isolation. Future work will focus on implementing the proposed method in full combining all four inputs (e.g. using a selection of machine learning techniques, like supervised learning, HMM, etc) and carrying out a large-scale evaluation to assess the system performance.

References

[1] Later Life in the United Kingdom, 2015, ageUK, http://www.ageuk.org.uk /professional-resources-home/policy/agenda-for-later-life/

[2] Rutherford, T. and Socio, A., 2012. Population ageing: statistics. *House of Commons library (Standard not. Retrieved Jan 2, 2013, from: www. parliament. uk/topics/PopulationArchive*.

[3] Lewin, D., Adshead, S., Glennon, B., Williamson, B., Moore, T., Damodaran, L. and Hansell, P., 2010. Assisted living technologies for older and disabled people in 2030. A final report to Ofcom. London. Plum Consulting

[4] Lewin, D., Adshead, S., Glennon, B., Williamson, B., Moore, T., Damodaran, L. and Hansell, P., 2010. Assisted living technologies for older and disabled people in 2030. A final report to Ofcom. London. Plum Consulting.

[5] Fleck, S. and Straßer, W., 2008. Smart camera based monitoring system and its application to assisted living. Proceedings of the IEEE, 96(10), pp.1698-1714.

[6] Fleck, S., Loy, R., Vollrath, C., Walter, F. and Straßer, W., 2007, September. SmartClassySurv-a smart camera network for distributed tracking and activity recognition and its application to assisted living. In Distributed Smart Cameras, 2007. ICDSC'07. First ACM/IEEE International Conference on (pp. 211-218). IEEE.

[7] Mundher, Z.A. and Zhong, J., A Real-Time Fall Detection System in Elderly Care Using Mobile Robot and Kinect Sensor.

[8] Galatas, G., Ferdous, S. and Makedon, F., 2013. Multi-modal person localization and emergency detection using the kinect. Int. J. Adv. Res. Artif. Intell, 2, pp.41-46.

[9] Patel, S. and Chauhan, Y., 2014. Heart attack detection and Medical attention using Motion Sensing Device-Kinect. International Journal of Scientific and Research Publications, p.468.

[10] Banerjee, T., Keller, J.M., Skubic, M. and Stone, E., 2014. Day or night activity recognition from video using fuzzy clustering techniques. Fuzzy Systems, IEEE Transactions on, 22(3), pp.483-493.

[11] Zhou, Z., Stone, E.E., Skubic, M., Keller, J. and He, Z., 2011, August. Nighttime in-home action monitoring for eldercare. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE (pp. 5299-5302). IEEE.

[12] Harvey, N., Zhou, Z., Keller, J.M., Rantz, M. and He, Z., 2009, September. Automated estimation of elder activity levels from anonymized video data. In Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE (pp. 7236-7239). IEEE.

[13] Haritaoglu, I., Harwood, D. and Davis, L.S., 2000. W 4: Real-time surveillance of people and their activities. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(8), pp.809-830.

[14] Guesgen, H.W., 2015, July. Towards a Theory of Space for Activity Recognition in Smart Environments Based on Rough Sets. In *Intelligent Environments (IE), 2015 International Conference on* (pp. 148-151). IEEE.

[15] Aztiria, A., Augusto, J.C., Izaguirre, A. and Cook, D., 2009. Learning accurate temporal relations from user actions in intelligent environments. In*3rd Symposium of Ubiquitous Computing and Ambient Intelligence 2008* (pp. 274-283). Springer Berlin Heidelberg.

[16] Pal, S. and Abhayaratne, C., 2015, September. Video-based activity level recognition for assisted living using motion features. In Proceedings of the 9th International Conference on Distributed Smart Camera (pp. 62-67). ACM

[17] Szegedy, C., Toshev, A. and Erhan, D., 2013. Deep neural networks for object detection. In *Advances in Neural Information Processing Systems*(pp. 2553-2561).